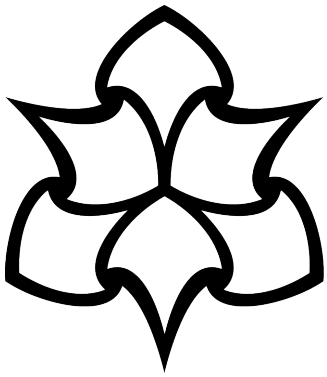

Breast Ultrasound Lesions Recognition: End-to-end Deep Learning Approaches

Moi Hoon Yap[a,*], **Manu Goyal**[a], **Fatima Osman**[b], **Robert Martí**[c], **Erika Denton**[d], **Arne Juette**[d], **Reyer Zwiggelaar**[e]

[a]Manchester Metropolitan University, Faculty of Science and Engineering, School of Computing, Mathematics and Digital Technology, Chester Street, Manchester, UK, M14 6TE

[b]Department of Computer Science, Sudan University of Science and Technology, Khartoum, Sudan.

[c]Computer Vision and Robotics Institute, University of Girona, Spain

[d]Nolfolk and Norwich University Hospital Foundation Trust, Norwich, UK

[e]Aberystwyth University, Department of Computer Science, Aberystwyth, SY23 3DB, UK

Abstract. Multi-stage processing of automated breast ultrasound lesions recognition is dependent on the performance of prior stages. To improve the current state of the art, we propose the use of end-to-end deep learning approaches using Fully Convolutional Networks (FCNs), namely FCN-AlexNet, FCN-32s, FCN-16s and FCN-8s for semantic segmentation of breast lesions. We use pre-trained models based on ImageNet and transfer learning to overcome the issue of data deficiency. We evaluate our results on two datasets, which consist of a total of 113 malignant and 356 benign lesions. To assess the performance, we conduct 5-fold cross validation using the following split: 70% for training data, 10% for validation data, and 20% testing data. The results showed that our proposed method performed better on benign lesions, with a top *Mean Dice* score of 0.7626 with FCN-16s, when compared to the malignant lesions with a top *Mean Dice* score of 0.5484 with FCN-8s. When considering the number of images with *Dice* score > 0.5, 89.6% of the benign lesions were successfully segmented and correctly recognised, while 60.6% of the malignant lesions were successfully segmented and correctly recognised. We conclude the paper by addressing the future challenges of the work.

Keywords: breast ultrasound, breast lesions recognition, fully convolutional network, semantic segmentation.

*Moi Hoon Yap, m.yap@mmu.ac.uk

1 Introduction

Breast cancer is the most common cancer in the UK [1], where one in eight women will be diagnosed with breast cancer in their lifetime and one person is diagnosed every 10 minutes [1]. Over recent years, there has been significant research into using different image modalities [2] and technical methods have been developed [3, 4] to aid early detection and diagnosis of the disease. These efforts have led to further research challenge and demand for robust computerised methods for cancer detection.

Two view mammography is known as the gold standard for breast cancer diagnosis [2]. However, ultrasound is the standard complementary modality to increase the accuracy of diagnosis.

Other alternatives include tomography and magnetic resonance, however, ultrasound is the most widely available option and widely used in clinical practice [5].

Conventional computerised methods in breast ultrasound cancer diagnosis comprised multiple stages, including pre-processing, detection of the region of interest (ROI), segmentation and classification [6–8]. These processes rely on hand-crafted features including descriptions in the spatial domain (texture information, shape and edge descriptors) and frequency domain. With the advancement of deep learning methods, we can detect and recognise objects without the need for hand-crafted features. This paper presents the limitation of the state of the art and conducts a feasibility study on the use of a deep learning approach as an end-to-end solution for fully automated breast lesion recognition in ultrasound images.

Two-Dimensional (2D) breast ultrasound lesion segmentation is a challenging task due to the speckle noise and being operator dependent. So far, image processing and conventional machine learning methods are deemed as preferable methods to segment the breast ultrasound lesions [9]. These are dependent on the human designed features such as texture descriptors [10, 11] and shape descriptors [7]. With the help of these extracted features, image processing algorithms [12] are used to locate and segment the lesions. Some of the state-of-the-art segmentation solutions consist of multiple stages [13, 14] - preprocessing or denoising stage, initial lesion detection stage to identify a region of interest [15] and segmentation [16]. Recently, Huang et al. [9] reviewed the breast ultrasound image segmentation solutions proposed in the past decade. In their study, they found that due to the ultrasound artifacts and to the lack of publicly available datasets for assessing the performance of the state-of-the-art algorithms, the breast ultrasound segmentation is still an open and challenging problem.

2 Related Work

This section summarises the state-of-the-art segmentation and classification approaches for breast ultrasound cancer analysis.

2.1 BUS Segmentation Approaches

Achieving an accurate segmentation in BUS images is considered to be a big challenge [17], because of the appearance of sonographic tumors [18, 19], the speckle noise, the low image contrast, and the local changes of image intensity [20]. Considering radiologist interaction within the segmentation process, it could have semi-automatic or fully automatic segmentation approaches [21].

Semi-automated segmentation approaches require an interaction with the user such as setting seeds, specifying an initial boundary or a region of interest (ROI). For instance, in [22], a computerized segmentation method for breast lesions on ultrasound images was proposed. First, a contrast-limited adaptive histogram equalization was applied. Then, in order to enhance lesion boundary and remove speckle noise, an anisotropic diffusion filter was applied, guided by texture descriptors derived from a set of Gabor filters. Further, the derived filtered image was multiplied by a constraint Gaussian function, to eliminate the distant pixels that do not belong to the lesion. To create potential lesion boundaries, a marker-controlled watershed transformation algorithm was applied. Finally, the lesion contour was determined by evaluating the average radial derivative function.

In order to segment ultrasonic breast lesions, Gao et.al. [18] proposed a variant of a normalized cut (NCut) algorithm that was based on homogeneous patches (HP-NCut) in 2012. Further, HPs were spread within the same tissue region, which is more reliable to distinguish the different tissues for better segmentation. Finally in the segmentation stage, they used the NCut framework

by considering the fuzzy distribution of textons within HPs as final image features. More recently, Prabhakar et.al. [23] developed algorithm for an automatic segmentation and classification of breast lesions from ultrasound images. s a pre-processing step, speckle noise was removed using the Tetrolet filter and, subsequently, active contour models based on statistical features were applied to obtain an automatic segmentation. For the classification of breast lesions, a total of 40 features were extracted from the images, such as textural, morphological and fractal features. Support Vector Machines (SVM) with a polynomial kernel for the combination of texture, optimal features were used to classify the lesions from BUS images.

Fully automatic segmentation needs no user intervention at all. In [24], instead of using a term-by-term translation of diagnostic rules on intensity and texture, a novel algorithm to achieve a comprehensive decision upon these rules was proposed. This was achieved by incorporating image over-segmentation and lesion detection in a pairwise conditional random field (CRF) model. In order to propagate object-level cues to segments, multiple detection hypotheses were used. Further, a unified classifier was trained based on the concatenated features. This algorithm could avoid the limitations of bottom-up segmentation, and capable to handle very complicated cases. In the same year, a novel algorithm was proposed [19], making no assumptions about lesions, in which a hierarchical over-segmentation framework was used for collecting heterogeneous features. Considering multiscale property, the superpixels were classified with their confidences nested into the bottom layer. An efficient CRF model was used for making the ultimate segmentation. Compared with other two different approaches, Hao et.al [19] algorithm was superior in performance, and was able to handle all kinds of tumors (benign and malignant).

In [25], two new concepts of neutrosophic subset and neutrosophic connectedness (neutro-connectedness) were defined to generalize the fuzzy subset and fuzzy connectedness. The newly

proposed neutro-connectedness models the inherent uncertainty and indeterminacy of the spatial topological properties of the image. The proposed method was applied to a BUS dataset with 131 cases, and its performance was evaluated using the similarity ratio, false positive ratio and average Hausdroff error. In comparison with the fuzzy connectedness segmentation method, the proposed method was more accurate and robust in segmenting tumors in BUS images.

2.2 BUS Classification Approaches

The majority of state-of-the-art methods are multi-stage. First to detect a lesion, i.e. where a lesion is localised on the image [26]. The localisation of a lesion can be done by manual annotation or using automated lesion detection approaches [6, 15]. Subsequently, next step is to identify the lesion type using feature descriptors. Amongst different proposed approaches considering solid mass classification, there are two main feature descriptors [27], i.e. echo texture [28] [11] and shape and margin features [29]. We present a couple of works on multi-stage machine learning methods. For a full review, please refer to Cheng et al. [26]. Liu et al. [30] proposed a novel breast classification system for Color Doppler flow imaging and B-Mode ultrasound. In order to obtain features from B-Mode ultrasound, many feature extraction methods were used to provide both the texture and geometric features. The first stage was an extraction of color Doppler features, which was achieved by applying blood flow velocity analysis to Doppler signals to extract several spectrum features. In addition, the authors proposed a velocity coherent vector method. Furthermore, using a support vector machine classifier, selected features were used to classify breast lesions into benign or malignant classes. They achieved an area under the ROC curve of 0.9455 when validated on 105 cases with 50 benign and 55 malignant. In the same year, Yap et al. [31] carried out a comprehensive analysis of the best feature descriptors and classifiers for breast ultrasound classification.

They experimented with 19 features (texture, shape and edge), 22 feature selection methods and ten classifiers. From their findings, the best combination was the feature set of 4 shape descriptors, 1 edge descriptor and 3 texture descriptors using a Radial Basis Function Network, with an area under the ROC curve of 0.948. In 2016, Yap and Yap [32] conducted study to evaluate the performance of machine learning on human delineation and computer method. They found that there were no significant differences for benign lesions but computer segmentation showed better accuracy for malignant lesion classification.

There is increasing interest in deep learning for medical imaging [33] and two research groups have been successful in using this in breast ultrasound. In 2016, Huynh et al. [34] proposed the use of a transfer learning approach for ultrasound breast images classification. The authors used 1125 cases and 2393 regions of interest for their experiment, where the ROIs were selected and labeled by the experts. To compare with the hand-crafted features, CNN was used to extract the features. When classify the CNN-extracted features with support vector machine on the recognition task of benign and malignant, they achieved an area under the ROC curve of 0.88. However, their solution was multi-stage and they did not share their dataset. In 2017, Yap et al. [35] demonstrated the use of deep learning for breast lesions detection, which outperformed the previous state-of-the-art image processing and conventional machine learning methods. They achieved an F-measure of 0.92 on breast lesions detection and made one of the dataset available for research purposes.

Recently, Yap et al. [36] demonstrated the practicality and feasibility of using a deep learning approach for automated semantic segmentation for BUS lesion recognition. However, they only performed one fold validation using one type of FCNs, i.e. FCN-AlexNet. This paper extends Yap et al. [36] to 5-fold cross validation on four types of FCNs, namely, FCN-AlexNet, FCN-32s, FCN-16s and FCN-8s. We are the first to implement semantic segmentation on BUS images.

3 Methodology

This section provides an overview of the breast ultrasound datasets, the preparation of the ground truth labeling, the proposed method and the type of performance metrics used to validate our results.

3.1 Datasets

To date, data deficiency in medical imaging analysis is a common problem. To form a larger dataset, we combined two datasets, which were the only two datasets made available for researchers. We provide a summary for each dataset and the details can be found in [35].

In 2001, a professional didactic media file for breast imaging specialists [37] was made available. It was obtained with B&K Medical Panther 2002 and B&K Medical Hawk 2102 US systems with an 8-12 MHz linear array transducer. Dataset A consists of 306 images from different cases with a mean image size of 377×396 pixels. From these images, 306 contained one or more lesions. Within the lesion images, 60 images presented malignant masses (as in Fig. 1 first row (a)) and 246 were benign lesions (as in Fig. 1 first row (b)). To obtain Dataset A, the user needs to purchase the didactic media file from Prapavesis et al. [37]. Yap et al. [35] named it as Dataset A in their description.

In 2012, the UDIAT Diagnostic Centre of the Parc Taulí Corporation, Sabadell (Spain) has collected Dataset B with a Siemens ACUSON Sequoia C512 system 17L5 HD linear array transducer (8.5 MHz). The dataset consists of 163 images from different women with a mean image size of 760×570 pixels, where the images presented one or more lesions. Within the 163 lesion images, 53 were malignant lesions (as in Fig. 1 first row (c)) and 110 with benign lesions (as in Fig. 1

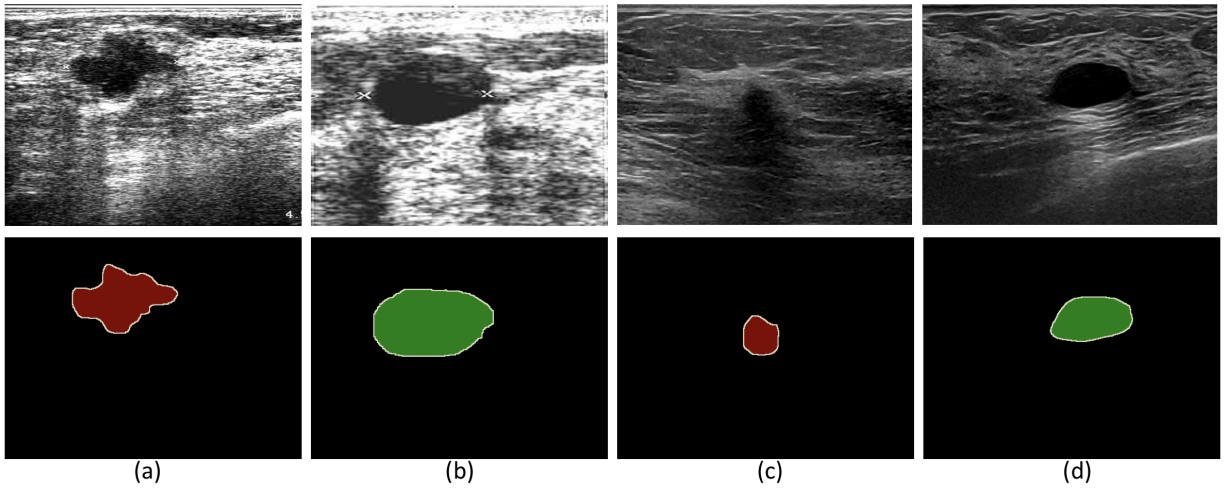

Fig 1 Illustration of some images from the datasets and its ground truth labeling in PASCAL-VOC format.(a) and (b) are images from Dataset A; (c) and (d) are images from Dataset B; and index 1 (RED) indicates malignant lesion and index 2 (GREEN) indicates benign lesion.

first row (d)). Dataset B and the respective delineation of the breast lesions are available online for research purposes, please refer to [35], where they named it as Dataset B in their description.

3.2 Ground Truth

Since deep learning models for semantic segmentation are widely evaluated for the PASCAL-VOC 2012 training and validation dataset, these trained models are tested for various performance metrics on the PASCAL-VOC 2012 test set [38, 39]. In the PASCAL-VOC 2012 dataset, the RGB images are used as input images. The dimensions of both input images and label images should be the same size [40]. Although the images used in training are not required to be the same size for deep learning models in segmentation tasks, all the images are required to be of same size due to the use of fully connected layers in these models. In the labelled image, every pixel value for each class is an index ranging from 0 to 255. In the PASCAL-VOC 2012 dataset, there are a total of 21 classes used so far, hence, 21 indexes are used for labelling the images. For breast ultrasound images, the format in digital media is generally grayscale. Hence, to make this compatible with the

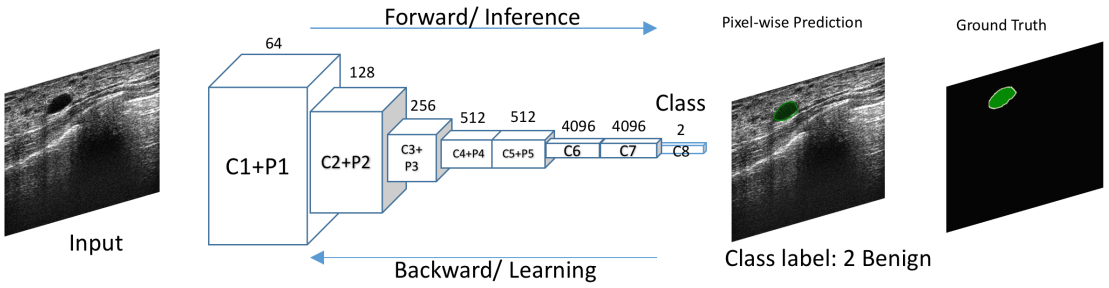

Fig 2 Overview of the semantic segmentation architecture.

pre-trained models and networks that are trained for PASCAL-VOC 2012 dataset (RGB images), we converted the grayscale images to RGB images with the help of channel conversion. The ground truths in binary masks format are converted into the 8-bit paletted label images. Fig. 1 illustrates the breast ultrasound images with the corresponding ground truth labeling in PASCAL-VOC format, where index 1 (RED) indicates malignant lesion and index 2 (GREEN) indicates benign lesion.

3.3 Deep Learning Framework

The deep learning methods proved its superiority over image processing methods and traditional machine learning in the detection of abnormalities in medical imaging of various modalities [35, 41]. There are two main types of tasks associated with medical imaging i.e. classification and semantic segmentation [42,43]. However, a known limitation of the classification is its inability to locate the abnormalities in medical imaging. Hence, semantic segmentation deep learning methods address this issues by classifying each pixel of the medical images rather than single prediction per image in the classification task. A popular group of deep learning methods for end-to-end semantic segmentation are fully convolutional networks (FCNs) [44].

FCN-AlexNet is a FCN version of the original AlexNet classification model with a few adjustments in the network layers for the segmentation task [44]. This network was originally used

for the classification of 1000 different objects of classes on the ImageNet dataset [45]. FCN-32s, FCN-16s, and FCN-8s are three models inspired by the VGG-16 based net which is a 16-layer CNN architecture that participated in the ImageNet Challenge 2014 and secured the first position in localization and second place in classification competition. All deep learning frameworks rely on feature extraction through the convolution layers, but classification networks throw away the spatial information in the fully connected layers. In contrast with classification network which ignores spatial information using fully connnected layers, FCN incorporates this information by replacing fully connected layers with convolution layers. Feature maps from those convolution layers are later used for classifying each pixel to get the semantic segmentation.

Transfer Learning is a procedure where a CNN is trained to learn features for a broad domain after which layers of the CNN are fine-tuned to learn features of a more specific domain. Under this setting, the features and the network parameters are transferred from the broad domain to the specific one depending on several factors such as size of the new dataset and similarity to the original dataset. The use of deep learning methods for semantic segmentation in medical imaging suffer from the problem of data deficiency, which can be overcome with the help of transfer learning approaches [41,42]. In this work, the pre-trained models on the ImageNet dataset which contains more than 1.5 millions images of 1000 classes was used for transfer learning [45]. The weights trained on ImageNet dataset are transferred for semantic segmentation of BUS with minor adjustments in the convolutionized fully connected layers [44]. We initialised the weights of convolutional layers from these pre-trained models rather than setting up the random weights for the limited medical datasets such as BUS dataset. Otherwise, it is very hard to converge the models based on the limited medical datasets. Hence, we fine-tuned these models by using pre-trained models and training on two classes i.e. benign and malignant in the BUS dataset as shown

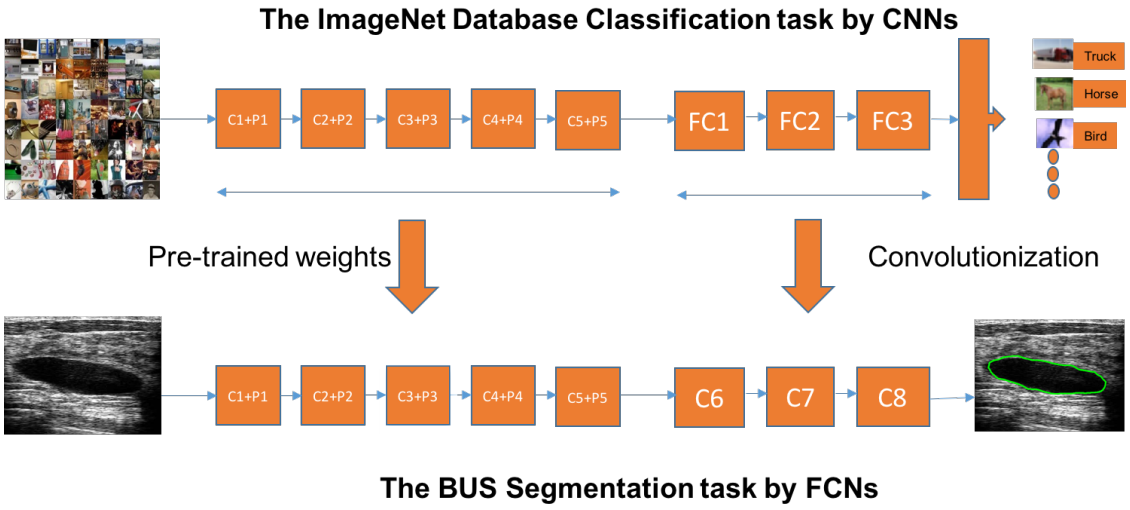

Fig 3 Transfer learning procedure of deep CNNs to obtain optimized weights initializations. Three fully connected layers of CNN were removed and replaced by three convolutional layers, making the pre-trained model fully convolutional.

in the Fig. 3.

The combination of Dataset A and Dataset B forms a larger dataset with a total of 113 malignant lesions and 356 benign lesions. We used the combined dataset to form better training and transfer learning to overcome the problem of data deficiency. We used DIGITS V5 which acts as a wrapper for the deep learning Caffe framework on the GPU machine of the following configuration: (1) Hardware: CPU - Intel i7-6700 @ 4.00Ghz, GPU - NVIDIA TITAN X 12Gb, RAM - 32Gb DDR5 (2) Deep Learning Framework: Caffe [46].

We assessed the performance of the model using 5-fold cross validation using the following split: 70% for training data, 10% for validation data, and 20% testing data. We trained the model using stochastic gradient descent with a learning rate of 0.0001, 60 epochs with a dropout rate of 33%. The number of epochs was kept at 60 as in [47] where convergence has already happened when we performed the empirical experiments. Fig. 2 illustrates the process of the end-to-end solution using semantic segmentation.

Table 1 Summary of the performances for different lesion types for four semantic segmentation methods in *Mean*. *SD* is standard deviation.

Lesion Type	Method	Sensitivity *Mean*±*SD*	Precision *Mean*±*SD*	Dice *Mean*±*SD*	MCC *Mean*±*SD*
Benign	FCN-AlexNet	0.8000±0.2404	0.7282±0.2191	0.7199±0.1964	0.7304±0.1762
	FCN-32s	0.8271±0.2250	0.7471±0.1923	0.7473±0.1896	0.7554±0.1689
	FCN-16s	**0.8374±0.2392**	0.7674±0.1953	**0.7626±0.2095**	**0.7733±0.1857**
	FCN-8s	0.8092±0.2683	**0.7940±0.1960**	0.7564±0.2373	0.7659±0.2172
Malignant	FCN-AlexNet	0.4708±0.3078	0.7599±0.2364	0.4894±0.2757	0.5080±0.2488
	FCN-32s	0.4492±0.2983	**0.7737±0.2925**	0.3267±0.2870	0.4001±0.2577
	FCN-16s	0.3790±0.2978	0.7481±0.2718	0.4212±0.2804	0.4616±0.2527
	FCN-8s	**0.5696±0.3350**	0.7044±0.2528	**0.5484±0.2785**	**0.5842±0.2358**

3.4 Evaluation criteria

Even though the method is an end-to-end solution, we evaluated the results using standard performance metrics from the literature. To measure the accuracy of the segmentation results, the *Dice Similarity Coefficient (Dice)* (henceforth *Dice*) [48, 49] was used. We report our findings in *Dice*, *Sensitivity*, *Precision* and *Matthew Correlation Coefficient (MCC)* [50] as our evaluation metrics.

4 Results and Discussion

Table 1 summarises the performance of our proposed methods on benign and malignant lesions. Overall, all the methods performed better on benign lesions, with a top *Dice* score of *0.7626*, compared to the malignant lesions with a top *Dice* score of *0.5484*. The results showed that the performance of the proposed method was dependent on the size of the dataset. In our datasets, we have more benign images (356) than malignant images (113). Overall, FCN-16s has the best performance in benign lesions recognition that achieved 0.8374 in *Sensitivity*, 0.7626 in *Dice Score* and 0.7733 in *MCC*. FCN-8s has the best *Precision* of 0.7940. For Malignant lesions, FCN-8s is the best method with 0.5696 in *Sensitivity*, 0.5484 in *Dice* and 0.5842 in *MCC*.

According to Everingham et al. [51], the results with *Dice* score > 0.5 is considered correct detection. Fig. 4 compares the performances of the proposed methods when considering the number

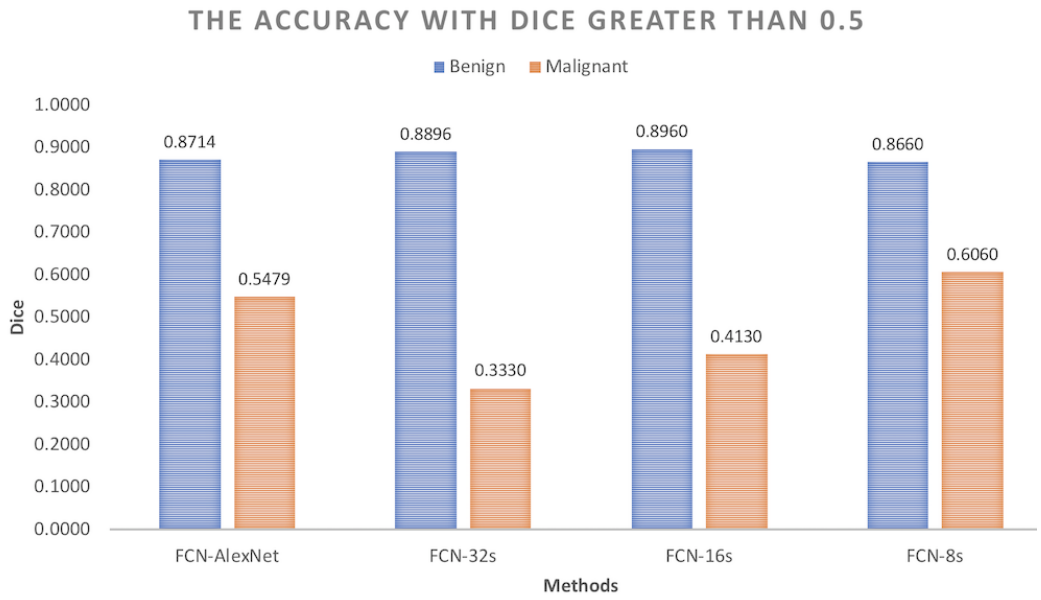

Fig 4 The accuracy of the proposed methods when considering the number of images with *Dice* score > 0.5.

of images with *Dice* score > 0.5. Overall, benign lesions had higher Dice score, with top accuracy of 0.8960 for FCN-16s. This implies that 89.6% of the benign lesions were successfully segmented and correctly recognised. The results were comparable across four different methods. For malignant lesions, the top accuracy is 0.6060 with FCN-8s, where only 60.6% of the malignant lesions were successfully segmented and correctly recognised. The worst performance in malignant lesions recognition was FCN-32s, where only 33.3% of the lesions was successfully segmented and recognised. The poor performances were due to data deficiency in malignant class, which is a common issue for deep learning approaches.

To further illustrate the results, we visually compared the segmented regions for the proposed methods. Four examples of the successful and failed cases for our experiment are illustrated in Fig. 5. The first row is a benign lesion, where the lesion is well-defined with clear boundaries. All the methods achieved high *Dice* score. Fig. 5 second row illustrates a malignant lesion with irregular boundaries and ill-defined shape. We observed that all the methods had classified the lesion to the

correct class. However, only FCN-16s managed to produce the closest segment when compared to the ground truth. The third row of Fig. 5 shows a benign lesion where all the methods failed to segment the lesion. This is due to the appearance of fibroadenoma are less hypo-echoic and poor image quality. The final row illustrates that even though the methods are able to segment the lesion, misclassification is an issue where FCN-AlexNet and FCN-32s have classified the hypo-echoic region as benign. FCN-8s are able to classify the lesion correctly however it also detected some benign regions within the lesion. Overall, the lesions with small area, ambiguity in the boundary and irregular shape are harder for semantic segmentation due to the lack of data to represent these categories.

5 Conclusion

The common problem in conventional machine learning are: 1) It is based on hand-crafted features; 2) In some cases, it requires human intervention where the radiologists has to select the ROI; and 3) It is multi-stage and there is dependency from one stage to the next. In this paper, the problem was solved by using a deep learning approach where we have shown four types of FCNs in designing a robust end-to-end solution for breast ultrasound lesions recognition.

Conventional methods classified the lesion into single type, but using semantic segmentation, we observed that it is not necessarily the case. In one lesion, as illustrated in Fig. 5 row 3 and row 4, it may have malignant tissue and benign tissue. This is an interesting finding for future research in understanding the tumour from both the computer vision and clinical perspectives.

This paper has provided a new insight for future research to by investigating four types of deep learning techniques. However, proposing an accurate end-to-end solution for breast ultrasound lesions recognition remains a challenge due to the lack of datasets to provide sufficient data repre-

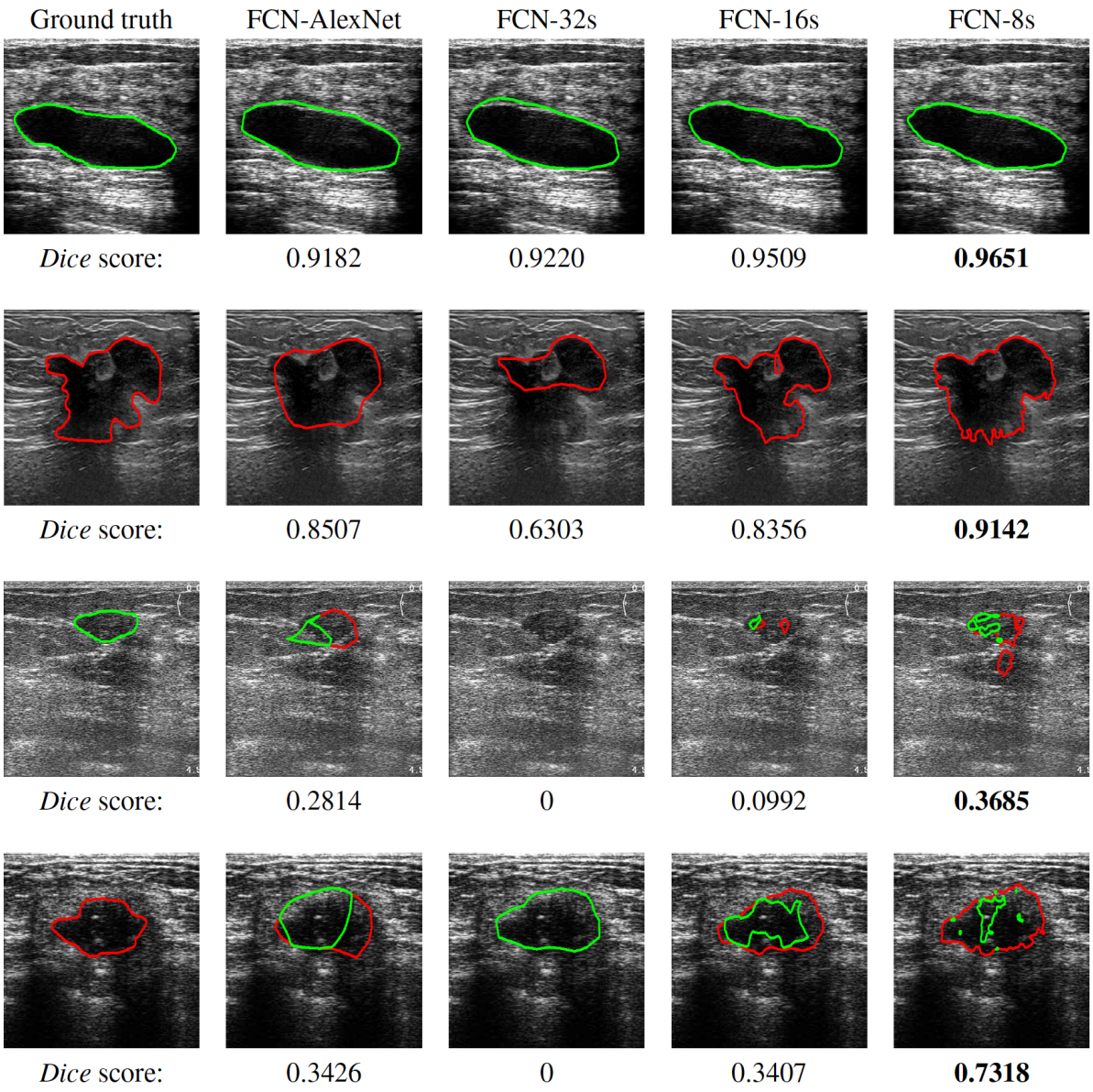

Fig 5 Visual comparison of the lesions segmentation and recognition with FCNs. The first column is the ground truth delineation, the second column is the proposed transfer learning FCN-AlexNet, the third column is the proposed transfer learning FCN-32s and the fourth column is the proposed transfer learning FCN-16s and the last column is the proposed transfer learning FCN-8s. The first and second rows showed the best case scenarios where the lesions were correctly segmented and classified. The third and fourth rows showed difficult cases where FCNs failed in those cases.

sentation. In the future, with the growth of big data and data sharing efforts, an end-to-end solution based on deep learning approach may find wide applications in breast ultrasound computer aided diagnosis.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

The authors would like to thank Prapavesis et al. (breast imaging specialists) [37] for providing Dataset A for this research.

References

1 "Breast cancer care: Facts and statistics 2017." Online. https://www.breastcancercare.org.uk/about-us/media/press-pack-breast-cancer-awareness-month/facts-statistics (2017).

2 W. Berg, L. Gutierrez, M. NessAiver, *et al.*, "Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer," *Radiology* **233**(3), 830–849 (2004).

3 M. H. Yap, A. G. Gale, and H. J. Scott, "Generic infrastructure for medical informatics (gimi): the development of a mammographic training system," in *International Workshop on Digital Mammography*, 577–584, Springer (2008).

4 M. H. Yap, E. Edirisinghe, and H. Bez, "Processed images in human perception: A case study in ultrasound breast imaging," *European Journal of Radiology* **73**(3), 682–687 (2010).

5 A. Stavros, C. Rapp, and S. Parker, *Breast Ultrasound*, 978-0397516247, LWW, 1 ed. (1995).

6 K. Drukker, M. L. Giger, C. J. Vyborny, *et al.*, "Computerized detection and classification of cancer on breast ultrasound," *Academic Radiology* **11**(5), 526–535 (2004).

7 M. H. Yap, E. Edirisinghe, and H. Bez, "A comparative study in ultrasound breast imaging classification," *Proc.SPIE* **7259**, 7259 – 7259 – 11 (2009).

8 J. Shan, H. Cheng, and Y. Wang, "A novel automatic seed point selection algorithm for breast ultrasound images," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, 1–4 (2008).

9 Q. Huang, Y. Luo, and Q. Zhang, "Breast ultrasound image segmentation: a survey.," *International journal of computer assisted radiology and surgery* **12**(3), 493–507 (2017).

10 A. Alvarenga, A. Infantosi, W. Pereira, *et al.*, "Assessing the combined performance of texture and morphological parameters in distinguishing breast tumors in ultrasound images," *Medical Physics* **39**(12), 7350–7358 (2012).

11 B. Liu, H. Cheng, J. Huang, *et al.*, "Fully automatic and segmentation-robust classification of breast tumors based on local texture analysis of ultrasound images," *Pattern Recognition* **43**(1), 280 – 298 (2010).

12 M. H. "Yap, E. A. Edirisinghe, and H. E. Bez, "Object boundary detection in ultrasound images," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, 53–53, IEEE (2006).

13 K. Drukker, N. P. Gruszauskas, C. A. Sennett, *et al.*, "Breast US computer-aided diagnosis workstation: Performance with a large clinical diagnostic population," *Radiology* **248**(2), 392–397 (2008).

14 J. Shan, H. Cheng, and Y. Wang, "Completely automated segmentation approach for breast

ultrasound images using multiple-domain features," *Ultrasound in Medicine and Biology* **38**(2), 262–275 (2012).

15 M. H. Yap, E. A. Edirisinghe, and H. E. Bez, "A novel algorithm for initial lesion detection in ultrasound breast images," *Journal of Applied Clinical Medical Physics* **9**(4), 181–199 (2008).

16 M. H. Yap, E. A. Edirisinghe, and H. E.Bez, "Fully automatic lesion boundary detection in ultrasound breast images," (2007).

17 H. Shao, Y. Zhang, M. Xian, *et al.*, "A saliency model for automated tumor detection in breast ultrasound images," 1424–1428 (2015).

18 L. Gao, W. Yang, Z. Liao, *et al.*, "Segmentation of ultrasonic breast tumors based on homogeneous patch," *Medical physics* **39**(6Part1), 3299–3318 (2012).

19 Z. Hao, Q. Wang, H. Ren, *et al.*, "Multiscale superpixel classification for tumor segmentation in breast ultrasound images," in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 2817–2820, IEEE (2012).

20 L. Gao, X. Liu, and W. Chen, "Phase-and gvf-based level set segmentation of ultrasonic breast tumors," *journal of applied Mathematics* **2012** (2012).

21 M. Xian, Y. Zhang, H. Cheng, *et al.*, "A benchmark for breast ultrasound image segmentation (busis)," *arXiv preprint arXiv:1801.03182* (2018).

22 W. Gomez, L. Leija, A. Alvarenga, *et al.*, "Computerized lesion segmentation of breast ultrasound based on marker-controlled watershed transformation," *Medical physics* **37**(1), 82–95 (2010).

348 23 T. Prabhakar and S. Poonguzhali, "Automatic detection and classification of benign and ma-
349 lignant lesions in breast ultrasound images using texture morphological and fractal features,"
350 in *Biomedical Engineering International Conference (BMEiCON), 2017 10th*, 1–5, IEEE
351 (2017).
352 24 Z. Hao, Q. Wang, Y. K. Seong, *et al.*, "Combining crf and multi-hypothesis detection for
353 accurate lesion segmentation in breast sonograms," in *International Conference on Medical*
354 *Image Computing and Computer-Assisted Intervention*, 504–511, Springer (2012).
355 25 M. Xian, H. Cheng, and Y. Zhang, "A fully automatic breast ultrasound image segmenta-
356 tion approach based on neutro-connectedness," in *Pattern Recognition (ICPR), 2014 22nd*
357 *International Conference on*, 2495–2500, IEEE (2014).
358 26 H. Cheng, J. Shan, W. Ju, *et al.*, "Automated breast cancer detection and classification using
359 ultrasound images: A survey," *Pattern Recognition* **43**(1), 299 – 317 (2010).
360 27 C. M. Sehgal, S. P. Weinstein, P. H. Arger, *et al.*, "A review of breast ultrasound," *Journal of*
361 *Mammary Gland Biology and Neoplasia* **11**(2), 113–123 (2006).
362 28 B. Sahiner, H.-P. Chan, M. A. Roubidoux, *et al.*, "Computerized characterization of breast
363 masses on three-dimensional ultrasound volumes," *Medical Physics* **31**(4), 744–754 (2004).
364 29 W. C. A. Pereira, A. V. Alvarenga, A. F. C. Infantosi, *et al.*, "A non-linear morphometric
365 feature selection approach for breast tumor contour from ultrasonic images," *Computers in*
366 *Biology and Medicine* **40**(11), 912–918 (2010).
367 30 Y. Liu, H. Cheng, J. Huang, *et al.*, "Computer aided diagnosis system for breast cancer based
368 on color doppler flow imaging," *Journal of Medical Systems* **36**(6), 3975–3982 (2012).
369 31 M. H. Yap, E. Edirisinghe, and H. Bez, "Computer aided detection and recognition of lesions

in ultrasound breast images," in *Innovations in Data Methodologies and Computational Algorithms for Medical Applications*, 125–152, IGI Global (2012).

32 M. H. Yap and C. H. Yap, "Breast ultrasound lesions classification: a performance evaluation between manual delineation and computer segmentation," in *SPIE Medical Imaging*, 978718–978718, International Society for Optics and Photonics (2016).

33 G. Carneiro, J. Nascimento, and A. P. Bradley, "Unregistered multiview mammogram analysis with pre-trained deep learning models," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 652–660, Springer (2015).

34 B. Huynh, K. Drukker, and M. Giger, "Computer-aided diagnosis of breast ultrasound images using transfer learning from deep convolutional neural networks," *Medical Physics* **43**(6), 3705–3705 (2016).

35 M. H. Yap, G. Pons, J. Mart, *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE Journal of Biomedical and Health Informatics* **22**, 1218–1226 (2018).

36 M. H. Yap, M. Goyal, F. Osman, *et al.*, "End-to-end breast ultrasound lesions recognition with a deep learning approach," in *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, **10578**, 1057819, International Society for Optics and Photonics (2018).

37 S. Prapavesis, B. Fornage, A. Palko, *et al.*, *Breast Ultrasound and US-Guided Interventional Techniques: A Multimedia Teaching File*, Thessaloniki, Greece (2003).

38 A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, *et al.*, "A review on deep learning techniques applied to semantic segmentation," *arXiv preprint arXiv:1704.06857* (2017).

39 M. Everingham, S. M. A. Eslami, L. Van Gool, *et al.*, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision* **111**, 98–136 (2015).

40 M. Thoma, "A survey of semantic segmentation," *arXiv preprint arXiv:1602.06541* (2016).

41 M. Goyal and M. H. Yap, "Multi-class semantic segmentation of skin lesions via fully convolutional networks," *arXiv preprint arXiv:1711.10449* (2017).

42 M. Goyal, M. H. Yap, N. D. Reeves, *et al.*, "Fully convolutional networks for diabetic foot ulcer segmentation," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 618–623 (2017).

43 M. Goyal, N. D. Reeves, A. K. Davison, *et al.*, "Dfunet: Convolutional neural networks for diabetic foot ulcer classification," *arXiv preprint arXiv:1711.10448* (2017).

44 J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440 (2015).

45 A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 1097–1105 (2012).

46 Y. Jia, E. Shelhamer, J. Donahue, *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 675–678, ACM (2014).

47 J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research* **12**(Jul), 2121–2159 (2011).

48 D. Zikic, Y. Ioannou, M. Brown, *et al.*, "Segmentation of brain tumor tissues with convolutional neural networks," *Proceedings MICCAI-BRATS* , 36–39 (2014).

49 S. Pereira, A. Pinto, V. Alves, *et al.*, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE Transactions on Medical Imaging* **35**(5), 1240–1251 (2016).

50 D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies* **2**, 37–63 (2011).

51 M. Everingham, L. Van Gool, C. K. Williams, *et al.*, "The pascal visual object classes (voc) challenge," *International journal of computer vision* **88**(2), 303–338 (2010).

Moi Hoon Yap is Reader (Associate Professor) in Computer Vision at the Manchester Metropolitan University and a Royal Society Industry Fellow with Image Metrics Ltd. She received her Ph.D. in Computer Science from Loughborough University in 2009. After her Ph.D., she worked as Postdoctoral Research Assistant in the Centre for Visual Computing at the University of Bradford. She serves as an Associate Editor for Journal of Open Research Software and reviewers for IEEE transactions/journals (Image Processing, Multimedia, Cybernetics, biomedical health, and informatics).

Manu Goyal is a Research Scholar at the Manchester Metropolitan University. He received his Master of Technology in Computer Science and Applications from Thapar University, India. His research expertise is in medical imaging analysis, computer vision, deep learning, wireless sensor networks and internet of things.

Fatima Mohamed Osman is currently Secretary of the Board of Trustees of the World Organization for Renaissance of Arabic Language, and a third year PhD student at Sudan University of

Science and Technology. She received her M.Sc. in Computer Science from University of Jordan, King Abdullah II School for Information Technology in Sep 2009. After her M.Sc. She served as a team member of Database Administration in IT Department at Zain Jordan Mobile Telecom (Jan 10 Dec 10). She worked as Teaching Assistant (Feb 2011 Nov 11) in the Department of Computer Science at the University of Africa.

Robert Martí is an associate professor at Computer Vision and Robotics Institute of the University Girona, Spain. He received his BS and MS degrees in Computer Science from the University of Girona in 1997 and 1999, respectively, and his PhD degree from the School of Information Systems at the University of East Anglia in 2003. He is the author of more than 30 international peer reviewed journal and 80 conference papers. His current research interests include machine learning and image registration applied to medical image analysis, computer aided diagnosis, and breast, prostate and brain imaging.

Erika Denton is a consultant radiologist in Norwich. Her appointment to the role of Associate Medical Director in 2016 is to provide leadership and strategic support to NNUHFT with specific responsibilities for working across the local STP footprint, with the UEA and to develop workforce and research strategies. In 2016 she was appointed to the role of Clinical Advisor in Imaging at NHS Improvement. This has included leading regional work with the Clinical Senate to review Interventional Radiology Services across the region.

Arne Juette is Consultant Radiologist and Ultrasound lead at Nolfolk and Norwich University Hospital.

Reyer Zwiggelaar received the Ir. degree in Applied Physics from the State University Gronin-

gen, Groningen, The Netherlands, in 1989, and the Ph.D. degree in Electronic and Electrical Engineering from University College London, London, UK, in 1993. He is currently a Professor in the Department of Computer Science, Aberystwyth University, UK. He is the author or co-author of more than 250 conference and journal papers. His current research interests include Medical Image Understanding, especially focusing on Mammographic and Prostate Data, Pattern Recognition, Statistical Methods, Texture-Based Segmentation, Biometrics, Manifold Learning and Feature-Detection Techniques. He is an Associate Editor of Pattern Recognition and Journal of Biomedical and Health Informatics (JBHI).

List of Figures

1 Illustration of some images from the datasets and its ground truth labeling in PASCAL-VOC format.(a) and (b) are images from Dataset A; (c) and (d) are images from Dataset B; and index 1 (RED) indicates malignant lesion and index 2 (GREEN) indicates benign lesion.

2 Overview of the semantic segmentation architecture.

3 Transfer learning procedure of deep CNNs to obtain optimized weights initializations. Three fully connected layers of CNN were removed and replaced by three convolutional layers, making the pre-trained model fully convolutional.

4 The accuracy of the proposed methods when considering the number of images with *Dice* score > 0.5.

5 Visual comparison of the lesions segmentation and recognition with FCNs. The first column is the ground truth delineation, the second column is the proposed transfer learning FCN-AlexNet, the third column is the proposed transfer learning FCN-32s and the fourth column is the proposed transfer learning FCN-16s and the last column is the proposed transfer learning FCN-8s. The first and second rows showed the best case scenarios where the lesions were correctly segmented and classified. The third and fourth rows showed difficult cases where FCNs failed in those cases.

List of Tables

1 Summary of the performances for different lesion types for four semantic segmentation methods in *Mean*. *SD* is standard deviation.